Distinguishing Between Stochastic Models of Heterogeneity and Contagion

by

Stanley S. Wasserman

Technical Report No. 332

November 1978

Department of Applied Statistics
University of Minnesota
St. Paul, Minnesota 55108

Support for this research was provided by the Social Science Research Council, and the University of Minnesota Graduate School and McMillan Travel Fund. This research was done while the author was visiting the Center for the Social Sciences at Columbia University, whose hospitality is gratefully acknowledged.

Abstract

Historically, the Poisson process has been the "benchmark" model for many social processes. When data from a particular social process fail to be adequately described by a Poisson distribution, a researcher may turn to generalized Poisson processes to more accurately model his or her empirical data. Two common generalized processes are the 1) heterogeneous Poisson process, in which the process rate is a random variable, and 2) contagious Poisson process, in which the process rate depends linearly on the current state of the process.

Paradoxically, both the heterogeneous and contagious processes yield the same theoretical distribution for the number of events that occur in an interval of time. Consequently, distinguishing between these two can be difficult. We discuss this situation, first reviewing the models and then giving strategies for choosing between them with empirical data.

Key words and phrases: Poisson process, compound Poisson process, contagious process, negative binomial distribution, event histories, embeddability waiting time distribution.

I. Introduction

For many years, the Poisson process has been the "benchmark" model for social and biological processes, a standard to which empirical observations are compared. As a first analytical stage, a researcher may model his or her process as a Poisson process. A thorough study of lack-of-fit of this model can be a valuable indicator of heterogeneity across objects or individuals and nonstationarity and/or nonhomogeneity in time. The residuals from the fitted model may allow the researcher to construct a more realistic model for the process, incorporating assumptions that negate the simple postulates of the purely random Poisson process.

Generalized Poisson processes are often used as "second stage" models. Two common processes are 1) the heterogeneous Poisson process, in which the rate of the process differs across individuals, and 2) the contagious Poisson process, in which the process rate depends on the current state of the process. These two models are quite different in their underlying structure, the first allowing individuals to evolve by distinct Poisson processes with rates independent of their current state, and the second specifying that individuals move identically through the states of the process but that this movement occurs at system dependent rates.

This paper studies an interesting paradox that arises when attempting to distinguish between models of heterogeneity and contagion. If we model individual heterogeneity by allowing the rate of the process to be a gamma random variable and if we build our contagious process so that the rate of the process is a linear function of the current state, then

we find that X(t), the number of events that occur in the time interval (0,t), is a negative binomial random variable under both models. Thus, with data only on the number of occurrences in a fixed time, we can not distinguish between these two quite different sources of nonrandomness.

We should note that there are two broad classes of contagious processes: infectious and addictive contagion. The distinction is based on whether the entire population of individuals or objects can or can not affect the probability of the occurence of an event for a single individual. In an infectious process, the probability of future events depends on the past history of all the individuals subject to the "infection"; e.g., in an epidemic, the probability of contracting a contagious disease certainly is a function of how many people in the population currently have the disease. In contrast, for an addictive or "self-infectious" process, the probability of future events depends only on the past history of the single individual being modelled. An example of an addictive process is hospitalization for schizophrenia. Future hospitalization depends only on the number and length of past hospitalization episodes for a specific patient and not on the event histories of all schizophrenics. Davis, Duncan, and Siverson (1978) make this distinction between addictive and infectious processes in a study of dynamic models for warfare, and Arbous and Kerrich (1951) note the applicability of self-infectious processes, but not infectious ones, to models of accident causation. The models that we discuss in this paper are all infectious stochastic processes, primarily because of the mathematical advantages of statistically independent individuals.

Prior to discussing our recommended strategies for discriminating between contagious and heterogeneous processes, we give a brief history of the development of these models and early attempts at discrimination. The heterogeneous Poisson process was introduced by Greenwood and Yule (1920). Eggenberger and Polya (1923) rediscovered the process and the associated negative binomial distribution, and presented a model for true infectious contagion. Lundberg (1940) and Feller (1943) were first to note the double nature of the negative binomial distribution arising from these two processes. Arbous and Kerrich (1950), Bates and Neyman (1952), and Lundberg (1940), applying these processes to accident statistics, discuss the problem of distinguishing between the models, but their proposed methods, which we review in later sections, do not make the best use of available data. Very little has been written on this problem in the statistics literature since these papers appeared in the early 1950's.

Sociologists have picked up the problem, and several papers have recently been published. Coleman (1965) introduces these models to social scientists and emphasizes the need for methods to distinguish between alternative causes of nonrandomness. Spilerman (1970, 1971) and Ritterband and Silberstein (1973) use the models to study racial disturbances and group disorders. Eaton (1974) and Tableson (1974) argue the merits of the correlational methods proposed by Arbous-Kerrich and Bates-Neyman for distinguishing between models, with the former demonstrating the usefulness of the methods on data on hospitalization of schizophrenia.

Taibleson, however, strongly states that these methods cannot "pin down" exactly the cause of the departure of such data from the random Poisson process. But, of course, his argument can be made for any statistical procedure used by social scientists to prove causation. It is our belief that these methods are useful and provide a first exploratory step in the discrimination procedure, prior to application of the embeddability and waiting time methods discussed in later sections of this paper. Very recently, Eaton and Fortin (1978), using the ideas of Quenouille (1949), present a third method for arriving at a negative binomial distribution for X(t), a compound logarithmic-Poisson distribution. This derivation is not based on a stochastic mechanism for X(t), and since the specification of postulates incorporating this logarithmic probabilistic component would be quite difficult, we shall not discuss this model.

In the next section, we present four models, the Poisson process, a heterogeneous Poisson process, and two contagious processes, giving their mathematical assumptions and origin. Following this exposition, we state the problem, discussing parameter estimation and the need for longitudinal data. In Section IV, we caution the researcher on selection of sampling strategies, demonstrating the estimation biases that occur for the contagious model parameters when individuals are sampled in a less than systematic fashion. We also propose a strategy for estimating the contagion parameter when individuals sampled do not have a common origin.

In Sections V, VI, and VII, the correlational methods of Arbous-Kerrich and Bates-Neyman for distinguishing between models are discussed and our embeddability and waiting time methods are presented. We show

how to estimate the underlying infinitesimal generator $\underset{\sim}{Q}$ by embedding the empirical observations on the process in a continuous time Markov chain in Section VI. By comparing the observed probability transition matrix with a theoretical estimate of it derived from both model assumptions and the data, we can compute goodness of fit statistics to test the fit of the data to the models. In Section VII, we give an alternative method for model discrimination based on the waiting times either between events or until the occurence of the Nth event. We derive the waiting time distributions for the heterogeneous and contagious processes, and discuss features of these previously unknown distributions.

II. Model Exposition

In this section we present the four models that are considered in later sections. All four are time-homogeneous pure birth Markov chains. Assume that we have a random sample of n individuals, and that for individual i, $X_i(t)$ is the number of "events" that occur in the time (0,t).

There are three assumptions common to the four models:

Postulate 1:

$X_i(t)$ is a Markov chain on the nonnegative integers, in continuous time, $t > 0$, for $i = 1,2,\ldots,n$.

Postulate 2:

$X_i(0) = 0$, a common origin for all individuals.

Postulate 3:

$$\Pr\{X_i(t+h) - X_i(t) > 1 | X_i(t) = x\} = o(h) \text{ as } h \to 0 \text{ for all } i \text{ and } x.$$

Postulates 1 and 3 specify the pure birth structure of the models. We consider the effect of relaxation of Postulate 2 in Section IV, where we assume the existence of a set of not necessarily equal starting times $\{t_{01}, t_{02}, \ldots, t_{0n}\}$, such that

$$X_1(t_{01}) = X_2(t_{02}) = \ldots = X_n(t_{0n}) = 0. \qquad (2.1)$$

Each of the four models has a set of postulates that specify the probability of the occurence of an event in the time interval (t,t+h). These model-specific postulates are numbered with a letter prefix for the appropriate model. The four models are: 1) The Poisson process (denoted by P), presented here simply as a basis of comparison for the other models; 2) A heterogeneous Poisson process (H) due to Greenwood-Yule, in which the birth-rate parameter λ is a gamma random variable, varying from individual to individual presumably because of some exogenous causation; 3) A contagious Poisson process of positive reinforcement (PR), in which the probability of an event occuring in (t,t+h) is an increasing linear function of the number of events occuring by time t. This model is due to Eggenberger-Polya, and is called an increasing linear growth model with immigration by Karlin and Taylor (1975); and 4) another contagious Poisson process, but of negative reinforcement, in which the infinitesimal probability of a "birth" is a decreasing linear function of X(t). Bates and Neyman (1952) discuss a fifth candidate, a contagious process with a

"learning" parameter that decreases the probabilities of event occurences with time, but since we are only considering time homogeneous models, we do not present this model here.

The model-specific postulates are as follows:

Poisson Process

Postulates:

$$P1 : \Pr\{X_i(t+h)-X_i(t) = 1 \mid X_i(t) = x\} = \lambda h + o(h)$$

$$P2 : \Pr\{X_i(t+h)-X_i(t) = 0 \mid X_i(t) = x\} = 1 - \lambda h + o(h)$$

Both postulates are true for all x, as $h \to 0$.

Heterogeneous Poisson Process

Postulates:

$$H1 : \Pr\{X_i(t+h)-X_i(t) = 1 \mid X_i(t) = x\} = \lambda_i h + o(h)$$

$$H2 : \Pr\{X_i(t+h)-X_i(t) = 0 \mid X_i(t) = x\} = 1 - \lambda_i h + o(h)$$

Both postulates are true for all x, as $h \to 0$.

$$H3 : \lambda_i \sim f_\Lambda(\lambda) = \frac{\beta}{\Gamma(\alpha)} (\beta\lambda)^{\alpha-1} e^{-\beta\lambda}, \qquad \lambda > 0.$$

The event rates are i.i.d. gamma random variables, with parameters $\alpha, \beta > 0$.

Contagious Positive Reinforcement Process

Postulates:

$$PR1 : \Pr\{X_i(t+h)-X_i(t) = 1 \mid X_i(t) = x\} = a + bx + o(h)$$

$$PR2 : \Pr\{X_i(t+h)-X_i(t) = 0 \mid X_i(t) = x\} = 1 - (a + bx) + o(h)$$

for $a,b > 0$ as $h \to 0$.

Contagious Negative Reinforcement Process

Postulates:

$$NR1 : \Pr\{X_i(t+h)-X_i(t) = 1 \mid X_i(t) = x\} = a - bx + o(h)$$

$$NR2 : \Pr\{X_i(t+h)-X_i(t) = 0 \mid X_i(t) = x\} = 1 - (a-bx) + o(h)$$

for $a,b > 0$, presumably with $a >> b$, as $h \to 0$.

III. Statement of the Problem

The problem arises when the Poisson process does not provide an adequate description for empirical observations $\{X_1(t) = x_1, X_2(t) = x_2, \ldots, X_n(t) = x_n\}$ on the number of events per individual occuring in $(0,t)$. The three latter models discussed in the previous section offer alternative explanations of the source or cause of the non-Poisson randomness in event occurences. As the following theorem states, we can only partially discriminate between these alternatives. The distributions (3.1)-(3.4) are common to all n individuals by previous postulates.

Theorem 1:

If $p_P(x)$, $p_H(x)$, $p_{PR}(x)$, and $p_{NR}(x)$ are probability mass functions, $Pr\{X(t) = x\}$, $x = 0,1,2,\ldots$, for the four models, then

$$p_P(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \tag{3.1}$$

$$p_H(x) = \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!}\left(\frac{\beta}{\beta+t}\right)^{\alpha}\left(\frac{t}{\beta+t}\right)^{x} \tag{3.2}$$

$$p_{PR}(x) = \frac{\Gamma\left(\frac{a}{b}+x\right)}{\Gamma\left(\frac{a}{b}\right)x!} e^{-at}\left(1-e^{-bt}\right)^{x} \tag{3.3}$$

$$p_{NR}(x) = \frac{\Gamma\left(\frac{a}{b}+1\right)}{\Gamma\left(\frac{a}{b}-x+1\right)x!} e^{-t(a-bx)}\left(1-e^{-bt}\right)^{x} \tag{3.4}$$

Proof:

The derivation of the Poisson distribution (3.1) is well known. To derive equation (3.2), first note that the conditional distribution is Poisson:

$$p_H(x|\lambda) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \tag{3.5}$$

To find the unconditional distribution, we use postulate H3, to get

$$p_H(x) = \int_0^{\infty} \frac{(\lambda t)^x e^{-\lambda t}}{x!} \frac{\beta}{\Gamma(\alpha)} (\beta\lambda)^{\alpha-1} e^{-\beta\lambda} d\lambda \tag{3.6}$$

which yields (3.2) upon integration.

The differential equations for the probability generating function (PGF) of X(t) under models PR and NR are

$$\frac{\partial G_{PR}(u,t)}{\partial t} = (u-1)\left[aG_{PR}(u,t) + bu \frac{\partial}{\partial u} G_{PR}(u,t)\right] \tag{3.7}$$

and

$$\frac{\partial G_{NR}(u,t)}{\partial t} = (u-1)\left[aG_{NR}(u,t) - bu \frac{\partial}{\partial u} G_{NR}(u,t)\right] \tag{3.8}$$

The solution to equation (3.7) is

$$G_{PR}(u,t) = \left[e^{bt} - u(e^{bt}-1)\right]^{-a/b} \tag{3.9}$$

which is the PGF of a negative binomial random variable with distribution (3.3). Differential equation (3.8) has solution

$$G_{NR}(u,t) = \left[e^{-bt} - u(e^{-bt} - 1)\right]^{a/b}$$

$$= \left[1 - (1 - e^{-bt}) + u(1 - e^{-bt})\right]^{a/b} \tag{3.10}$$

the PGF for a binomial random variable B(N,p), with $N = \frac{a}{b}$ and $p = 1 - e^{-bt}$, verifying (3.4). Q.E.D.

Thus, with data on occurences of events in single intervals, if the data are either Poisson or binomial, we can confidently choose between a Poisson process and a negative reinforcement process as a stochastic model. However, if these data appear to be negative binomial, our discrimination task is only slightly simplified: we can rule out models P and NR, but can not choose between models H and PR because of their common negative binomial distributions (3.2) and (3.3) for X(t). The remainder of this paper is devoted to this discrimination task.

The key to distinguishing between H and PR is longitudinal data. At the very least, data on the number of events in the intervals (s,s+t) and (u,u+v), $s \geq 0$ and $u \geq s+t$ are required. Coleman (1965, page 301) realizes this necessity, but also cautions the researcher in interpreting deviation from "a straight Poisson as being due to contagion rather than heterogeneity," since contagion may give the appearance of heterogeneity and vice versa. In an ongoing contagious process, individuals may have had time to differentiate and develop different event rates, giving evidence of spurious heterogeneity. And in a heterogeneous process, data may appear to have been generated by contagion because of positive correlations of events occuring in consecutive intervals, long thought to be good indicators of true contagion. But, as we show in Section V, both models yield a positive correlation of $[X(s+t)-X(s)]$ and $[X(u+v)-X(u)]$.

IV. Sampling the Process at Discrete Points in Time

Prior to presentation of methods useful for distinguishing between models, we discuss a problem that may occur when sampling individuals at discrete points in time to obtain data on the occurences of events. Suppose we sample the population of n individuals at times t and t+s and record

$$X_{1i} = X_i(t) - X_i(0)$$

$$X_{2i} = X_i(t+s) - X_i(s)$$

for all individuals, $s \geq t$, the number of events that occured during the time intervals (0,t) and (s,t+s).

Now, because of the true randomness of event occurences under the P and H models, p_P and p_H, distributions defined in equations (3.1) and (3.2), are common to all individuals, and more importantly, depend only on the length of the time interval. Hence,

$$p_P(x_{1i}) = p_P(x_{2i})$$

and (4.1)

$$p_H(x_{1i}) = p_H(x_{2i})$$

for i = 1,2,...,n. But this is not true for models PR and NR, because of the dependence of the distribution on the number of events that have occured at the beginning of the time interval. We can easily derive, for all i,

$$p_{PR}(x_{2i}) = \frac{\Gamma\left(\frac{a}{b}+x_{2i}\right)}{\Gamma\left(\frac{a}{b}\right)x_{2i}!}\left[e^{bs}(e^{bt}-1)+1\right]^{-\left(\frac{a}{b}+x_{2i}\right)}\left[e^{bs}(e^{bt}-1)\right]^{x_{2i}} \qquad (4.2)$$

and

$$p_{NR}(x_{2i}) = \frac{\Gamma\left(\frac{a}{b}+x_{2i}\right)}{\Gamma\left(\frac{a}{b}-x_{2i}\right)x_{2i}!}\left[1-e^{-bs}(1-e^{-bt})\right]^{\frac{a}{b}-x_{2i}}\left[e^{-bs}(1-e^{-bt})\right]^{x_{2i}} \qquad (4.3)$$

The distribution of X_{2i} is still either negative binomial (under PR) or binomial (under NR) but is quite different than that of X_{1i}. The differences are best seen by examination of the first two moments, given in Table 1. Note how, under both models, the moments have e^{bs} or e^{-bs} factors.

Positive Reinforcement (PR) Model

	X(t)	X(s+t)-X(s)
Expected Value	$\frac{a}{b}(e^{bt}-1)$	$\frac{a}{b}e^{bs}(e^{bt}-1)$
Variance	$\frac{a}{b}e^{bt}(e^{bt}-1)$	$\frac{a}{b}e^{bs}(e^{bt}-1)[e^{bs}(e^{bt}-1)+1]$

Negative Reinforcement (NR) Model

	X(t)	X(s+t)-X(s)
Expected Value	$\frac{a}{b}(1-e^{-bt})$	$\frac{a}{b}e^{-bs}(1-e^{-bt})$
Variance	$\frac{a}{b}e^{-bt}(1-e^{-bt})$	$\frac{a}{b}e^{-bs}(1-e^{-bt})[1-e^{-bs}(1-e^{-bt})]$

Table 1. Moments for X_1 and X_2 under reinforcement models.

Because of this dependence in the distributions on the beginning time of the interval, it is important that all individuals are sampled at exactly the same time in the evolution of the process. If we record events for a given time period, say a month, for a group of individuals, but allow the specific month to differ from individual to individual, then $X_i(s + 1 \text{ month}) - X_i(s)$, $i = 1,2,\ldots,n$, will not be identically distributed random variables.

Also note, that if Postulate 2, a common origin, is not true, then the problem also arises. Even if the sampled interval is identical across the population, with respect to both length and time of sampling, then the number of events that occur in the interval has a different distribution from individual to individual. Coleman (1965, page 301) analyzes date on purchases of phonograph records in a one month period assuming

that the positive reinforcement model and Postulate 2 hold. But it is very doubtful that the individuals studied have a common origin, and consequently, these data can not be analyzed as independent and identically distributed observations on a negative binomial random variable. This naivete´ is probably quite common.

The underlying reason for the nonequality of (4.2) and (3.3), and (4.3) and (3.4) is due to the nature of the reinforcement process: the probability of the next event increases with time. Consequently, the distribution of $X(s + t) - X(s)$ depends on s. If there is no common origin, let t_{0i} be the origin for the ith individual, i = 1, 2,...,n, such that $X_i(t_{0i}) = 0$. What follows is an exploratory strategy for estimating the parameters in this situation.

First define $t_0^* = \max_i\{t_{0i}\}$. Call the individual who has t_0^* as its origin, k. Then Postulates PR1 and PR2 (and NR1 and NR2) can be revised to reflect the inequality amongst the $\{t_{0i}\}$:

Postulate PR1*:

$$\text{a)}\quad P\{X_i(t_0^* + h) - X_i(t_0^*) = 1 \mid X_i(t_0^*) = x_i\} = a_i + o(h)$$

$$\text{b)}\quad P\{X_i(t + h) - X_i(t) = 1 \mid X_i(t) - X_i(t_0^*) = x,\ t > t_0^*\} = a_i + bx + o(h)$$

as $h \to 0$, where $a_i = a + bx_i$, $i \neq k$, and of course $a_k = a$.

Postulate PR2*:

$$a)\quad P\{X_i(t_0^* + h) - X_i(t_0^*) = 0 \mid X_i(t_0^*) = x_i\} =$$

$$1 - a_i + o(h)$$

$$b)\quad P\{X_i(t + h) - X_i(t) = 0 \mid X_i(t) - X_i(t_0^*) = x,\ t > t_0^*\} =$$

$$1 - (a_i + bx) + o(h)$$

as $h \to 0$, where a_i is defined as above.

Postulates NR1* and NR2* are similar, with the sign of b reversed. For $t > t_0^*$, we have a heterogeneous contagious process, with differential a_i.

Suppose that we have a priori knowledge of the existence of a mean and variance of a_i, say μ_a and σ_a^2. If so, we can get rough estimates of b using data $X_{1i} = (X_i(s + t) - X_i(s))$ and $X_{2i} = (X_i(u + t) - X_i(u))$, $i = 1, 2, \ldots, n$, $u \geq s + t$, as the following theorem states.

Theorem 2:

Assume the individuals in the population do not have a common origin, in violation of Postulate 2, and that Postulates PR1* and PR2* or NR1* and NR2* hold. Then, assuming the existence of a mean μ_a and variance σ_a^2 of the $\{a_i\}$,

$$\tilde{b}_{1PR} = \frac{1}{s - u} \ln\ (\bar{X}_1/\bar{X}_2)$$

$$\tilde{b}_{2PR} = \frac{1}{2(s - u)} \ln \left[(S_1^2 - \bar{X}_1)/(S_2^2 - \bar{X}_2)\right] \qquad (4.4)$$

are method-of-moment estimators for the contagion parameter b under model PR, where $\bar{X}_j = \frac{1}{n}\Sigma X_{ji}$ and $S_j^2 = \frac{1}{n}\Sigma(X_{ji} - \bar{X}_j)^2$, $j = 1,2$.

Under model NR,

$$\tilde{b}_{1NR} = \frac{1}{u - s}\ln(\bar{X}_1/\bar{X}_2)$$

$$\tilde{b}_{2NR} = \frac{1}{2(u - s)}\ln\left[(S_1^2 - \bar{X}_1)/(S_2^2 - \bar{X}_2)\right] \tag{4.5}$$

Proof:

We know, from Table 1, that for all i,

$$E(X_i(s + t) - X_i(s)|a) = E(X_{1i}|a) = \frac{a}{b}e^{bs}(e^{bt} - 1) \tag{4.6}$$

and

$$Var(X_i(s + t) - X_i(s)|a) = Var(X_{1i}|a) = \frac{a}{b}e^{bs}(e^{bt}-1)\left[e^{bs}(e^{bt} - 1) + 1\right] \tag{4.7}$$

with the PR model. Therefore,

$$E(X_{1i}) = E_a\left[E(X_{1i}|a)\right] = \frac{\mu_a}{b}e^{bs}(e^{bt} - 1) \tag{4.8}$$

and

$$\mathrm{Var}(X_{1i}) = \mathrm{Var}_a\left[E(X_{1i}|a)\right] + E_a\left[\mathrm{Var}(X_{1i}|a)\right] =$$

$$\frac{\sigma_a^2}{b^2} e^{2bs}(e^{bt} - 1)^2 + \frac{\mu_a}{b^2} e^{bs}(e^{bt} - 1)\left[e^{bs}(e^{bt} - 1) + 1\right] \qquad (4.9)$$

The mean and variance of X_{2i} are identical to (4.8) and (4.9), with u replacing s. Thus, we have the ratios

$$\frac{E(X_{1i})}{E(X_{2i})} = e^{b(s - u)}$$

and (4.10)

$$\frac{\mathrm{Var}(X_{1i}) - E(X_{1i})}{\mathrm{Var}(X_{2i}) - E(X_{2i})} = e^{2b(s - u)}$$

Substituting $\bar{X}_1/\bar{X}_2$ and $(S_1^2 - \bar{X}_1)/(S_2^2 - \bar{X}_2)$ for the ratios in (4.10) yields the moment estimators (4.4). Identical calculations give estimators (4.5) for the NR model. Q.E.D.

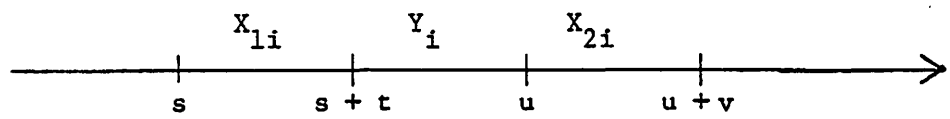
Method-of-moment estimators are more widely used with negative binomial data than maximum likelihood (ML) estimators since the ML equations have no closed-form solution (Shenton and Myers, 1963, or Johnson and Kotz, 1969.)

Throughout this section, we have assumed that all individuals in some population have been sampled. If we have data for some of the individuals in the population, then the preceding calculations are still true as long as the data are from a simple random sample. However, if the probablity that an individual is in the sample at time s is proportional to the number of events that have occured, $X_i(s)$, then we have a length-biased sample of the population. The distribution of $X_i(s + t) - X_i(s)$ is no longer negative binomial, but can be computed using the mathematical results of Zelen (1974).

V. Longitudinal Data

As mentioned in an earlier section, the key to distinguishing between models of heterogeneity and positive reinforcement is longitudinal data. Such data are usually the number of events that occur in a finite collection of time intervals for each and every individual. As discussed in the previous section, it is important that there be a common origin for all individuals. In this section, we give bivariate distributions and correlations of $X_{1i} = X_i(s + t) - X_i(s)$ and $X_{2i} = X_i(u + v) - X_i(u)$. Most of these calculations can be found in Arbous and Kerrich (1951).

Let the time interval $(s,u+v), s \geq 0$, be divided into three subintervals $(s,s + t)$, $(s + t,u)$, and $(u,u + v)$, and define X_{1i}, Y_i, and X_{2i} as the number of events that occur in the three intervals, for individuals $i = 1, 2,\ldots,n$. The situation is depicted in Figure 1. The marginal distributions of X_{1i}, Y_i, X_{2i}, and $X_i = X_{1i} + X_{2i} + Y_i$ are given in equations (3.1) - (3.4) or equations (4.2) - (4.3).



$$X_{1i} = X_i(s + t) - X_i(s)$$

$$X_{2i} = X_i(u + v) - X_i(u)$$

$$Y_i = X_i(u) - X_i(s + t)$$

Figure 1. Longitudinal data.

Theorem 3 gives the bivariate distributions of X_{1i} and X_{2i} for the models H, PR, and NR. With the pure Poisson process (model P), X_{1i}, Y_i and X_{2i} are jointly independent. Because of the zero correlations, it is easy to decide with longitudinal data whether model P is operating; consequently, we will focus our attention on models H, PR, and NR. The following distributions (5.1) - (5.3) are common to all n individuals.

Theorem 3:

The joint distributions of X_1 and X_2 are

$$P_H(x_1,x_2) = \frac{\Gamma(\alpha + x_1 + x_2)}{\Gamma(\alpha)x_1!x_2!} \frac{t^{x_1} v^{x_2} \beta^{\alpha}}{(\beta + t)^{\alpha + x_1 + x_2}} \tag{5.1}$$

$$P_{PR}(x_1,x_2) = \frac{\Gamma(\frac{a}{b} + x_1 + x_2)}{\Gamma(\frac{a}{b})x_1!x_2!} Q^{-\frac{a}{b}} \left(\frac{P_1}{Q}\right)^{x_1} \left(\frac{P_2}{Q}\right)^{x_2} \tag{5.2}$$

$$P_{NR}(x_1,x_2) = \frac{\Gamma(\frac{a}{b} + 1)}{\Gamma(\frac{a}{b} + 1 - x_1 - x_2)x_1!x_2!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{\frac{a}{b} - x_1 - x_2} \tag{5.3}$$

for $x_1, x_2 = 0,1,2,\ldots$, where

$$P_1 = e^{b(s+t)}(1 - e^{-bt})$$

$$P_2 = e^{b(u+v)}(1 - e^{-bv})$$

$$Q = 1 + P_1 + P_2$$

$$\pi_1 = e^{-bs}(1 - e^{-bt})$$

$$\pi_2 = e^{-bu}(1 - e^{-bv})$$

$$\pi_3 = 1 - \pi_1 - \pi_2$$

Proof:

The bivariate compound Poisson distribution (5.1) is derived by Arbous and Kerrich (1951, page 415). The negative trinomial distribution (5.2) is found by noting that X_1 and X_2 are negative binomial random variables, with a common parameter $N = \frac{a}{b}$, and hence have a joint negative trinomial distribution. The same reasoning is true for the binomial random variables X_1 and X_2 under NR, giving the derivation of the trinomial distribution (5.3). Q.E.D.

As a corollary to the theorem, we give the correlations between X_1 and X_2, without proof:

Corollary 1:

The correlations of X_1 and X_2, for all individuals i, are:

$$\rho_H(x_1,x_2) = \left[\left(1 + \frac{\beta}{t}\right)\left(1 + \frac{\beta}{v}\right)\right]^{-\frac{1}{2}} \tag{5.4}$$

$$\rho_{PR}(x_1,x_2) = \left[\left(1 + \frac{1}{P_1}\right)\left(1 + \frac{1}{P_2}\right)\right]^{-\frac{1}{2}} \tag{5.5}$$

$$\rho_{NR}(x_1,x_2) = -\left[\left(1 - \frac{1}{\pi_1}\right)\left(1 - \frac{1}{\pi_2}\right)\right]^{-\frac{1}{2}} \tag{5.6}$$

Note that correlations under model NR, unlike H and PR, are always negative, a fact that can be very useful in verifying the existence of negative reinforcement.

Arbous and Kerrich (1951) and Bates and Neyman (1952) recommend the use of correlations (5.4) and (5.5) to distinguish between models of heterogeneity and positive reinforcement. These correlations theoretically differ as long as $P_1 \neq t/\beta$ and $P_2 \neq v/\beta$, an unlikely event. The utility of comparing theoretical and empirical correlations depends on the stability of the operational time of the processes (Feller, 1966, page 178) as pointed out by Taibleson (1974). But operational times <u>are</u> stable by assumption, and if not, modelling the empirical observations is a difficult task. Also, Eaton and Fortin (1978) show that $\rho_H(x_1,x_2)$ does not depend on the length of the intervening true interval (s + t,u), unlike $\rho_{PR}(x_1,x_2)$, and that $\rho_{PR}(x_1,y) < \rho_{PR}(y,x_2)$, if t = v, where Y = X(u) - X(s + t), unlike correlations under model H.

With long individual event histories, we could easily check whether or not each individual evolved by an independent Poisson process, thoroughly testing the heterogeneity assumptions, presumably even postulate H3. But rarely are such data collected. Comparing empirical sample correlations to their theoretical equivalents is a good exploratory strategy, but often does not allow the researcher to unequivocally state that either model H or model PR is operating. In the next section we show how longitudinal observations can be used to find the specific model that has generated the data, whether it be H, PR, or some different model. These methods also allow χ^2 tests for goodness-of-fit.

VI. Distinguishing between models

Suppose we have sampled the individuals at times s and s + t, and recorded $X_i(s)$ and $X_i(s + t)$, i = 1, 2, ..., n. With these longitudinal data, we can estimate the infinitesimal generator $\underset{\sim}{Q}$ of the underlying Markov process, and determine whether a heterogeneity or contagion model is operating. We make no apriori assumptions on the nature of the infinitesimal generator -- we only assume a continuous-time stationary Markov chain as a model. Each of our four models has an associated class of generators Q, unique to each model. By comparing the empirically calculated infinitesimal generator with the four distinct classes of $\underset{\sim}{Q}$ matrices we can determine which, if any, model is operating. We simply check for the embeddability of the longitudinal data into one of our four classes of continuous time Markov chains.

In this section, we first give a brief review of theory for continuous time Markov chains, and compute the four classes of infinitesimal generators. We then discuss the computation of estimates of $\underset{\sim}{Q}$ and the comparison of these estimates to their theoretical counterparts, and lastly, testing for goodness-of-fit to one of the four models.

Throughout, we assume longitudinal data from only two sampled time points, s and s + t. With observations from more than two time points, additional estimates of $\underset{\sim}{Q}$ can be computed. Singer and Spilerman (1974, 1976) give guidelines for utilizing more than one estimate of $\underset{\sim}{Q}$. Singer (1977) discusses the general problem of recovering infinitesimal generators using incomplete or partial individual event histories. With more data than X(s) and X(s + t), the analytic strategy is identical to that given here, with the

exception that a researcher must now check more than one $\underset{\sim}{Q}$ estimate for embeddability.

A stochastic process with a finite number of states has transition probabilities derived from the solution of the system of ordinary differential equations

$$\frac{d\underset{\sim}{P}(t)}{dt} = \underset{\sim}{Q}\underset{\sim}{P}(t) \tag{6.1}$$

where $\underset{\sim}{P}(t)$ and $\underset{\sim}{Q}$ are square N x N matrices. A well-known result is that if $\underset{\sim}{Q}$ has the infinitesimal generator structure

$$\begin{aligned} &q_{ij} \geq 0, \text{ for } i \neq j \\ &q_{ii} \leq 0, \text{ for all } i \\ &\sum_{j=i}^{N} q_{ij} = 0, \text{ for all } i \end{aligned} \tag{6.2}$$

then the matrix functions $\underset{\sim}{P}(t)$, the solutions of (6.1) for $t > 0$, are the <u>transition matrices</u> for continuous-time stationary Markov chains. The elements of $\underset{\sim}{P}(t)$, have the conditional probability interpretation

$$p_{ij}(t) = P\{X(t) = j \mid X(0) = i\}. \tag{6.3}$$

$\underset{\sim}{Q}$ is called the intensity matrix or <u>infinitesimal generator</u> of the process. The solution to (6.1) is

$$\underset{\sim}{P}(t) = e^{t\underset{\sim}{Q}}, \; t > 0 \tag{6.4}$$

where $e^{\cdot}$ is the matrix exponential function.

Each of our four models has an associated class of intensity matrices satisfying (6.2) that can be found from the model postulates. These classes, denoted by $\mathcal{Q}_P$, $\mathcal{Q}_H$, $\mathcal{Q}_{PR}$, and $\mathcal{Q}_{NR}$, are specified by the following theorem.

Theorem 4:

The model-specific classes of intensity matrices are defined as follows:

a) $\mathcal{Q}_P$ = set of all intensity matrices $\underset{\sim}{Q}$ with

$$q_{i,i+1} = -q_{ii} = \lambda$$

$$q_{ij} = 0,\ j - i \neq 0, 1.$$

b) $\mathcal{Q}_H$ = set of all intensity matrices $\underset{\sim}{Q}$ with

$$q_{i,i+1} = -q_{ii} = \alpha/\beta,$$

$$q_{ij} = 0,\ j - i \neq 0, 1.$$

c) $\mathcal{Q}_{PR}$ = set of all intensity matrices $\underset{\sim}{Q}$ with

$$q_{i,i+1} = -q_{ii} = a + bi,$$

$$q_{ij} = 0,\ j - i \neq 0, 1.$$

d) $\mathcal{Q}_{NR}$ = set of all intensity matrices $\underset{\sim}{Q}$ with

$$q_{i,i+1} = -q_{ii} = a - bi,$$

$$q_{ij} = 0,\ j - i \neq 0, 1.$$

Proof:

The classes $\mathcal{Q}_P$, $\mathcal{Q}_{PR}$, and $\mathcal{Q}_{NR}$ follow directly from their model-specific postulates given in Section II. To derive $\mathcal{Q}_H$, first note that the individual-level process has a $\underset{\sim}{Q}$ matrix in the class $\mathcal{Q}_P$; however, the population-level process is a gamma mixture of these constant diagonal/super-diagonal intensity matrices. Thus, $\underset{\sim}{Q} \in \mathcal{Q}_H$ is of the form

$$\underset{\sim}{Q} = \int \lambda(\underset{\sim}{M} - \underset{\sim}{I}) f_{\Lambda}(\lambda) d\lambda \qquad (6.5)$$

where $\underset{\sim}{M}$ is the super-diagonal unity matrix, $M_{i,i+1} = 1$, $M_{ij} = 0$ for $j - i \neq 1$. The matrix integration (6.5) is done element-by-element to yield the class $\mathcal{Q}_H$ defined above.

Q.E.D.

Note that $\mathcal{Q}_H$, $\mathcal{Q}_{PR}$, and $\mathcal{Q}_{NR}$ are mutually exclusive classes of intensity matrices, so that distinguishing between the models can be accomplished once the underlying $\underset{\sim}{Q}$ has been estimated from the data. Also note that $\mathcal{Q}_P = \mathcal{Q}_H$, set of matrices with constant diagonal and super-diagonal entries, with the two constants summing to zero, so that distinguishing between models P and H cannot be done with this strategy; however, these two models are so different with respect to the correlations of occurences in consecutive time intervals (see Section V) that distinguishing between them should not be difficult. Indeed, model P predicts zero correlations, while H, positive correlations.

To estimate $\underset{\sim}{Q}$, we form the maximum likelihood estimate of $\underset{\sim}{P}(t)$ (Anderson and Goodman, 1957). If we define the matrix of transitions $T = (t_{kl})$ as

t_{kl} = number of individuals with $X_i(s) = k$ and $X_i(s + t) = l$

then $\hat{\underset{\sim}{P}}(t)$ with elements $(\hat{P}_{kl}(t) = t_{kl}/t_{k\cdot})$ is the maximum likelihood estimate of $\underset{\sim}{P}(t)$, the probablity transition matrix for the process defined in equation (6.4).

Thus

$$\hat{\underset{\sim}{Q}} = \frac{1}{t} \log \hat{\underset{\sim}{P}}(t) \tag{6.6}$$

is an estimate of $\underset{\sim}{Q}$. $\hat{\underset{\sim}{Q}}$ will be unique if $\hat{\underset{\sim}{P}}(t)$ has distinct, positive, real eigenvalues (Singer and Spilerman, 1976, page 41). Since $\hat{\underset{\sim}{P}}(t)$ is a triangular matrix, its eigenvalues are its diagonal elements $\{\hat{p}_{ii}(t)\}$, which are positive and real, and are distinct with large probablity.

Because of sampling and/or measurement error, $\hat{\underset{\sim}{Q}}$ may not be a diagonal/super-diagonal matrix. Hence, we must find a $\hat{\underset{\sim}{Q}}$ in one of the classes Q, such that

$$\tilde{\underset{\sim}{Q}} = \min_{\tilde{\underset{\sim}{Q}} \in Q} \left\| \tilde{\underset{\sim}{Q}} - \hat{\underset{\sim}{Q}} \right\| ; \tag{6.6}$$

i.e., find the "closest" $\tilde{\underset{\sim}{Q}}$ to $\hat{\underset{\sim}{Q}}$, such that $\tilde{\underset{\sim}{Q}} \in Q$. $\tilde{\underset{\sim}{Q}}$ may be found by least squares. Four $\tilde{\underset{\sim}{Q}}$ matrices will be generated, one for each model.

Then, we can compare the observed $\hat{\underset{\sim}{P}}(t)$ with the four matrices

$$\tilde{\underset{\sim}{P}}(t) = e^{t\tilde{\underset{\sim}{Q}}} \tag{6.7}$$

and compute chi-squared statistics

$$X^2 = \sum_{i,j} \frac{(\tilde{p}_{ij} - \hat{p}_{ij})^2}{\tilde{p}_{ij}} \tag{6.8}$$

to test the hypotheses that each of the four models are operating. This approach also has the advantage that we need not accept any of these hypotheses. We may find that a model other than one of the four is operating.

VII. Alternative method of distinguishing

An alternative way of choosing between models of heterogeneity and contagion utilizes data on waiting times until the occurence of the Nth event. Under the Poisson process model, waiting times are gamma random variables. In this section, we derive previously unknown distributions of the waiting time until the Nth event assuming the heterogeneity model and the reinforcement models. Bates and Neyman (1952), conditioning on the occurence of an event in the interval $(t, t + 1)$, found the distribution of $\tau\varepsilon(0,1)$, the waiting time until the occurence. Here, we give unconditional waiting time distributions, which should prove more useful in practice, and method-of-moment estimators of the distribution parameters.

Let W_N be a random variable denoting the waiting time until the Nth event occurs, $W_N \geq 0$, with density $g_N(w)$. These densities are common to all individuals, so we drop the dependence on subscript i. The density function g_{NH} for the heterogeneity model is a multiple of a modified Bessel function of the second kind, while the function g_{NR} for the reinforcement models is a linear combination of exponential densities.

Theorem 5:

a) Assuming that the postulates of the heterogeneity model are true, then the distribution of W_N, $N = 1, 2, \ldots$, is

$$g_{NH}(w) = \frac{2\beta}{\Gamma(N)\Gamma(\alpha)}(w\beta)^{\frac{1}{2}(\alpha + N) - 1}K_{\alpha - N}(2\sqrt{w\beta}) \qquad (7.1)$$

where $\alpha > N$, $\beta > 0$, $w > 0$, and $K_{\alpha - N}(\cdot)$ is a modified Bessel function of the second kind of order $\alpha - N$.

b) Assuming that the postulates of either the positive or negative reinforcement model are true, then the distribution of W_1 is exponential:

$$g_{1R}(w) = \frac{1}{a} e^{-\frac{w}{a}}, \quad a > 0, \ w \geq 0. \tag{7.2}$$

The distribution of W_N, $N = 2, 3, \ldots$, is

$$g_{NR}(w) = \frac{1}{(N-1)!\beta^{N-1}} \left[\sum_{j=0}^{N-1} (-1)^{j+N-1} \binom{N-1}{j} (a \pm jb)^{N-2} e^{-\frac{w}{a \pm jb}} \right] \tag{7.3}$$

where $\alpha, \beta > 0$, $w \geq 0$. The + sign is correct for model PR, the - sign for model NR.

Proof:

With model H, density g_{NH} is a gamma mixture of a gamma:

$$W_N/\lambda \sim \text{Gamma}\ (N, \lambda^{-1})$$

$$\lambda \sim \text{Gamma}\ (\alpha, \beta)$$

Hence,

$$g_{NH}(w) = \frac{\beta^{\alpha} w^{N-1}}{\Gamma(N)\Gamma(\alpha)} \int_0^{\infty} \lambda^{\alpha-N-1} e^{-(\lambda\beta + w/\lambda)} d\lambda. \tag{7.4}$$

Noting that

$$\int_0^{\infty} \lambda^{\alpha-N-1} e^{-(\lambda\beta+w/\lambda)} d\lambda = 2\left(\frac{w}{\beta}\right)^{\frac{\alpha-N}{2}} K_{\alpha-N}(2\sqrt{\beta w}), \tag{7.5}$$

one can derive (7.1) with some algebra. For part b, distribution (7.2) is easily derived. When $N \geq 2$, let

$$W_N = Z_0 + Z_1 + \ldots + Z_{N-1}$$

where Z_i has exponential density

$$f_i(z) = (a \pm ib)\, e^{-\frac{z}{a \pm ib}}, \quad z \geq 0. \tag{7.6}$$

The density g_{NR} can be found inductively by successive convolutions.

Q.E.D

The moments of W_N assuming heterogeneity are found using the integral

$$\int_0^\infty t^{u-1} K_v(t)dt = 2^{u-2}\Gamma(\tfrac{u+v}{2})\Gamma(\tfrac{u-v}{2}), \tag{7.7}$$

and assuming reinforcement using the moment generating function

$$M_N(t) = \prod_{i=0}^{N-1} \left[1 - (a \pm ibt)\right]^{-1} \tag{7.8}$$

Corollary 2:

The mean and variance of the waiting time until the Nth event are

$$\mu_{NH} = N\,\alpha/\beta$$
$$\sigma^2_{NH} = N\,\alpha/\beta^2(1+N+\alpha) \tag{7.9}$$

or

$$\mu_{NR} = N\left[a \pm \frac{N-1}{2}b\right]$$
$$\sigma^2_{NR} = N\left[a^2 \pm (N-1)ab + \frac{(N-1)(2N-1)}{6}b^2\right] \tag{7.10}$$

assuming the heterogeneity or reinforcement assumptions, respectively.

Maximum likelihood estimates of the parameters α, β, or a, b must be found iteratively. However method-of-moment estimates of the parameters are easy to derive.

Suppose we have n observed waiting times until the Nth occurence, $W_{1N}, W_{2N}, \ldots, W_{nN}$. Let $\bar{W}_N = \frac{1}{n}\Sigma W_{iN}$, $S^2_N = \frac{1}{n}\Sigma(W_{iN} - \bar{W}_N)^2$ be the sample mean and variance of the waiting time. Then method-of-moment estimates are

$$\tilde{\alpha} = \frac{\bar{W}_N^2}{NS_N^2 - \bar{W}_N^2}\,(N+1)$$

$$\tilde{\beta} = \frac{NS_N^2 - \bar{W}_N^2}{\bar{W}_N N(N+1)} \qquad (7.11)$$

$$\tilde{b} = \frac{2\sqrt{3}}{N}\sqrt{\frac{NS_N^2 - \bar{W}^2}{(N-1)(N+1)}}$$

$$\tilde{a} = \frac{\bar{W}_N}{N} \pm \frac{N-1}{2}\,\tilde{b} \qquad (7.12)$$

These estimates are defined for $NS_N^2 > \bar{W}_N^2$, and $\bar{W}_N > \frac{N(N-1)}{2}\,\tilde{b}$.

One can take the empirical waiting time data categorize them, and compare the observed frequencies to expected frequencies obtained from these distributions, using parameter estimates (7.11) and (7.12). Then, χ^2 goodness-of-fit statistics, with n-3 degrees of freedom, will aid in the distinguishing task.

VIII. References

Anderson, T.W. and L.A. Goodman (1957) "Statistical inference about Markov chains." Annals of Mathematical Statistics, volume 28, pages 89-110.

Arbous, A.G. and J.E. Kerrich (1951) "Accident statistics and the concept of accident-proneness." Biometrics, volume 7, pages 340-432.

Bates, G.E. and J. Neyman (1952) "Contributions to the theory of accident proneness." University of California Publications in Statistics, volume 1, pages 215-75.

Coleman, J.S. (1965) Introduction to Mathematical Sociology. Glencoe, Illinois: The Free Press.

Davis, W.W., G.T. Duncan, and R.M. Siverson (1978) "The dynamics of warfare: 1816-1965." American Journal of Political Science, to appear.

Eaton, W.W. (1974) "Mental hospitalization as a reinforcement process." American Sociological Review, volume 39, pages 242-60.

Eaton, W.W. and A. Fortin (1978) "A third interpretation for the generating process of the negative binomial distribution." American Sociological Review, volume 43, pages 264-7.

Eggenberger, F. and G. Polya (1923) "Uber die Statistik verketteter Vorgange." Zeitschrift für Angewandte Mathematik und Mechanik, volume 1, pages 279-89.

Feller, W. (1943) "On a general class of contagious distributions." Annals of Mathematical Statistics, volume 14, pages 389-400.

Feller, W. (1966) An Introduction to Probability Theory and its Applications, volume 2. New York: John Wiley and Sons.

Greenwood, M. and G.U. Yule (1920) "An inquiry into the nature of frequency distribution representative of multiple happenings with particular reference to the occurence of multiple attacks of disease or of repeated accidents." Journal of the Royal Statistical Society, volume 83, pages 255-79.

Johnson, N.L. and S. Kotz (1969) Distributions in Statistics: Discrete Distributions. New York: John Wiley & Sons.

Karlin, S. and H.M. Taylor (1975) A First Course in Stochastic Processes, Second Edition. New York: Academic Press.

Lundberg, O. (1940) On Random Processes and their Applications to Sickness and Accident Statistics. Uppsala: Almquist and Wiksells.

Quenouille, M.H. (1949) "A relation between the logarithmic, poisson, and negative binomial series." Biometrics, volume 5, pages 162-4.

Ritterband, P. and R. Silberstein (1973) "Group disorders in the public schools." American Sociological Review, volume 38, pages 461-7.

Shenton, L.R. and R. Myers (1963) "Comments on estimation for the negative binomial distribution." Proceedings of the International Symposium on Discrete Distributions, Montreal, pages 241-62.

Singer, B. (1977) "Individual histories as the focus of analysis in longitudinal surveys", paper presented at the SSRC Conference on the National Longitudinal Surveys, Washington, D.C. October 1977.

Singer, B. and S. Spilerman (1974) "Social mobility models for heterogeneous populations." Sociological Methodology 1973-74, edited by H.L. Costner. San Francisco: Jossey-Bass.

Singer, B. and S. Spilerman (1976) "Representation of social processes by Markov models." American Journal of Sociology, volume 82, pages 1-54.

Spilerman, S. (1970) "The causes of racial disturbances: A comparison of alternative explanations." American Sociological Review, volume 35, pages 627-49.

Spilerman, S. (1971) "The causes of racial disturbances: Tests of an explanation." American Sociological Review, volume 36, pages 427-42.

Taibleson, M.H. (1974) "Distinguishing between contagion, heterogeneity, and randomness in stochastic models." American Sociological Review, volume 39, pages 877-80.

Zelen, M. (1974) "Problems in cell kinetics and the early detection of disease," in Reliability and Biometry. Philadelphia: SIAM, pages 701-26.